ViB-OPD: Visual Evidence Mining and Boundary-Balanced Video Distillation

Anonymous authors
Paper under double-blind review

Abstract

Existing approaches commonly use supervised fine-tuning (SFT) or reinforcement learning (RL) to develop video understanding and reasoning; however, SFT is constrained by the quality and coverage of fixed demonstrations, while RL often relies on sparse final-answer rewards that provide little direct guidance for intermediate reasoning. On-policy distillation (OPD) offers an alternative by providing teacher distributional supervision along trajectories generated by the current student. Yet applying OPD to video reasoning leaves two issues unresolved: teacher distributions can be limited by incomplete visual evidence, and token-wise aggregation can make supervision allocation depend on expression length. We propose ViB-OPD, comprising Visual Evidence Mining (VEM) and Boundary-Balanced OPD (BB-OPD). A stronger frozen teacher uses supplementary training-time visual context to guide the student without expanding its inference input. VEM combines global observation, question-conditioned localization, and optional local reinspection to organize candidate evidence into a separate contact sheet, enriching the teacher's visual basis. BB-OPD initializes semantic groups and combines local boundary perturbation with within-segment averaging and equal segment weights to reduce the influence of expression length and a single segmentation on supervision allocation. It preserves the original student trajectory and token-level distillation objective. Evaluation on six public benchmarks and across model families demonstrates training gains.

1 Introduction

Video understanding requires models not only to recognize entities and actions in individual frames, but also to connect state changes, event order, and evidence distributed over time (Fu et al., 2024; Li et al., 2023; Liu et al., 2024). Video language models have substantially advanced these capabilities (Zhang et al., 2023; Lin et al., 2023; Li et al., 2024; Bai et al., 2025b;a), yet their reasoning is still commonly trained with supervised fine-tuning (SFT) or reinforcement learning (RL) (Feng et al., 2025; Wang et al., 2025). SFT is constrained by the quality and coverage of fixed demonstrations, whereas RL typically derives sparse feedback from final-answer correctness and does not directly constrain intermediate reasoning. On-policy distillation (OPD) offers a complementary route: a stronger teacher provides dense distributional supervision on prefixes generated by the current student (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024). Directly applying OPD to video, however, does not necessarily provide effective supervision: the teacher's token-wise distribution depends on the visual evidence it observes, whereas the effectiveness of that supervision depends on how these signals are aggregated along the trajectory. Existing video OPD commonly assumes by default that the teacher and student share the same sampled frames and that all tokens are weighted equally, but neither setting is necessarily well suited to video.

What a teacher can supervise is bounded by the evidence it can see. Evidence decisive to a video question may occur only in a brief local event (Wu et al., 2024; Zhou et al., 2024); if limited sampling misses that event, even a stronger reasoner can only supervise from incomplete evidence. Uniformly expanding visual coverage increases input and computation (Zhang et al., 2024) and can carry this extra cost into deployment. Learning with

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

privileged information instead allows a training-time teacher to exploit supplementary information unavailable to the deployed student (Vapnik & Izmailov, 2015; Lopez-Paz et al., 2015; Ye et al., 2026). This lets us place question-relevant supplementary observations on the teacher side alone: local evidence is made salient during training, while the student retains its original fixed visual input at inference.

Even with adequate evidence, weighting every token equally is already an allocation of supervision rather than a neutral default. A visual observation, temporal relation, or intermediate judgment may play a comparable reasoning role yet receive a different total weight simply because it is verbalized with more tokens; prior work on process supervision and token reweighting likewise suggests that intermediate steps and token positions should not be treated uniformly (Uesato et al., 2022; Lightman et al., 2024; Wang et al., 2023; Wu et al., 2026; Jin et al., 2026). Grouping a trajectory into semantic units reduces this length dependence, but semantic boundaries are inherently non-unique: when supervision depends on one exact partition, a boundary error redistributes weight across neighboring tokens and writes the incidental cuts of a single segmentation into a persistent allocation. We therefore use semantic units as the basis for aggregation and introduce controlled variation in their boundary neighborhoods, allocating supervision across multiple local partitions while preserving the original trajectory and token-level distillation objective.

To this end we introduce ViB-OPD. Its **Visual Evidence Mining (VEM)** combines global observation, question-conditioned localization, and optional local reinspection into a supplementary contact sheet visible to the teacher, without transferring construction-stage analysis text. Its **Boundary-Balanced OPD (BB-OPD)** organizes the student trajectory into semantic units and combines local boundary perturbation, within-segment averaging, and equal segment weights, preventing expression length alone from determining process supervision while reducing dependence on a single exact segmentation. After training, only the student is retained, with no additional visual input or inference module.

We evaluate ViB-OPD on six public video benchmarks and two model families, Qwen3-VL and Qwen3.5, with controlled comparisons of visual conditioning, loss aggregation, and perturbation scale. On all 2,700 Video-MME questions, the model achieves 63.52%, exceeding token-averaged OPD with the same VEM-conditioned teacher by 1.00 percentage points overall and 2.11 points on the Long subset. With Qwen3.5-9B, Video-MME improves from 63.59% to 65.67%. Controlled aggregation comparisons further show that semantic initialization with local boundary perturbation outperforms all four fixed segmentation schemes, supporting a design that adjusts supervision within boundary neighborhoods anchored by semantic structure.

Our contributions are:

- **Visual Evidence Mining (VEM).** We export global observation, question-conditioned localization, and local reinspection as supplementary evidence sheets, enriching teacher visual context without transferring construction-stage reasoning text or adding student inference modules.
- **Boundary-Balanced OPD (BB-OPD).** We combine semantic grouping, local perturbation, and segment balancing to control segment-level weights and neighboring tokens' supervision, preserving same-trajectory token-level reverse KL.
- **Multi-benchmark evaluation and aggregation analysis.** Six public benchmarks and two model families validate the training gain; component and aggregation comparisons separate visual conditioning from supervision allocation and demonstrate the effectiveness of combining semantic grouping with local boundary perturbation.

2 Related Work

2.1 Video Understanding and Visually Privileged Learning

Video language models improve video understanding through visual alignment, temporal modeling, and longer context (Lin et al., 2023; Cheng et al., 2024; Li et al., 2024; Zhang et al., 2024), while video post-training further develops reasoning through reward design (Feng

et al., 2025; Wang et al., 2025; Qin et al., 2026). Privileged-information learning lets teachers use information available only during training (Vapnik & Izmailov, 2015; Lopez-Paz et al., 2015): OPCD guides student trajectories with enhanced context (Ye et al., 2026), and Token-Budget Distillation uses a full-visual-token teacher to train a compressed student (Guo et al., 2026). VEM exports question-conditioned observation and localization results as a separate evidence sheet, enriching teacher conditioning while retaining the student's inference input. This differs from passing construction-stage textual reasoning chains to the student.

2.2 On-Policy Distillation and Process Supervision

GKD matches teacher distributions on student-generated trajectories, while MiniLLM and DistiLLM study distribution-matching objectives and trajectory sampling (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024); VideoEcho-R1 rewrites teacher reasoning chains to reduce policy differences (Lee et al., 2026). Process supervision constructs rewards from step annotations or continuation success rates (Lightman et al., 2024; Wang et al., 2023; Wu et al., 2026), whereas Entropy-Aware OPD adjusts KL objectives at high-entropy positions (Jin et al., 2026). BB-OPD retains on-policy sampling and token-level reverse KL, concentrating its changes on semantic grouping, boundary placement, and segment aggregation without rewriting reasoning or predicting segment importance.

3 Method

3.1 Overview

ViB-OPD comprises Visual Evidence Mining (VEM) and Boundary-Balanced OPD (BB-OPD), which address the visual basis of teacher supervision and its allocation along student trajectories, respectively. Given video V, question q, and options o, the student π_θ receives only a uniformly sampled 32-frame global view V_G. The frozen teacher π_T additionally receives a 16-frame evidence contact sheet C_E constructed offline by VEM:

$$x_S = (V_G, q, o), \qquad x_T = (V_G, C_E, q, o). \tag{1}$$

Our main configuration uses a Qwen3-VL-8B student and a Qwen3-VL-32B teacher with a shared tokenizer and vocabulary; θ denotes the trainable student parameters. The student generates an on-policy trajectory, the teacher supplies distributions at the same text prefixes, and BB-OPD organizes this position-wise supervision. In Figure 1, VEM constructs evidence (A), while BB-OPD shows same-trajectory scoring and loss aggregation in one panel (B). Deployment retains only the student and x_S, without the evidence builder, teacher, or segmenter.

3.2 Visual Evidence Mining (VEM)

VEM supplies question-relevant details that global sampling may miss without expanding student inputs. To establish both when events occur and what is visible locally, the same frozen 32B model performs global observation, question-conditioned localization, and optional reinspection, organizing candidate evidence as an image. These stages establish context, index time, and verify details, respectively.

Global observation: establishing event context. The builder observes video clips and summarizes them into a temporally associated overview. The prompt requests salient visible content, explicitly retaining details such as on-screen text. This overview contextualizes subsequent localization, relating local actions to surrounding scenes and the question's reference events rather than judging isolated frames.

Temporal localization: mapping questions to intervals. Using the overview, question, and options, the builder represents candidate clips by start and end times and issues local inspection instructions when needed. This maps question-level event cues to readable video regions while retaining temporal relationships. Localization receives neither reference answers nor student rollouts; its intervals guide subsequent observation and frame sampling.

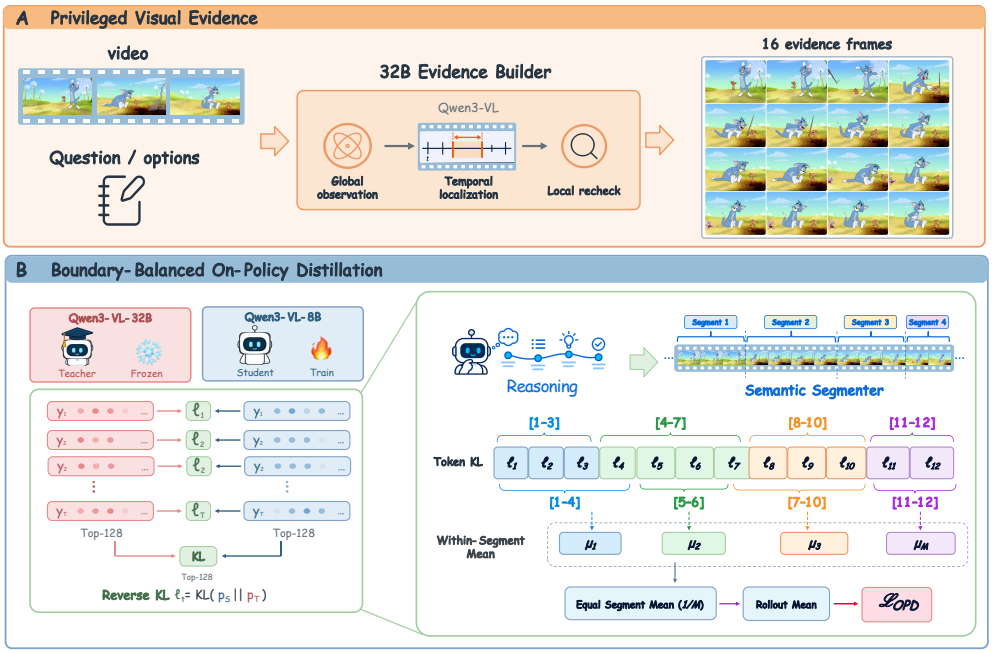

Figure 1: Training framework of ViB-OPD. **(A)** VEM mines supplementary visual evidence. **(B)** BB-OPD computes reverse KL along the same student trajectory, then updates the student through local boundary perturbation and segment-balanced aggregation.

Local reinspection: recovering visible details. When observations are insufficient, the builder reinspects candidate intervals according to the instructions, combining whole-clip and finer-subclip descriptions. Fine-grained prompts explicitly attend to visible text and overlays; other inspection targets depend on the question. Reinspection counts vary across examples. Export prioritizes reinspected intervals, otherwise using available localization intervals.

Evidence organization and text isolation. We distribute 16 sampling slots approximately equally across valid intervals, sample uniformly while prioritizing positions absent from the global view, and arrange frames chronologically into a timestamped 4×4 contact sheet C_E. Supplied as a separate image alongside V_G, it combines global context with local candidate evidence. Observation descriptions and localization analyses select intervals but are not exported as reasoning text or option judgments. The online teacher receives only the video, sheet, question, and options in Eq. 1. Since VEM increases teacher visual budget, its overall benefit cannot directly be attributed to selection quality.

3.3 Boundary-Balanced OPD (BB-OPD)

BB-OPD organizes token-level on-policy supervision into reasoning units with uncertain boundaries, while preserving the sampled student trajectory and position-wise teacher–student alignment. Its premise is that a visual observation, temporal relation, or intermediate inference often spans multiple tokens, whereas ordinary token averaging gives a reasoning step total weight proportional to its expression length. BB-OPD retains standard position-wise distributional supervision, constructs locally variable reasoning units, and then allocates supervision through a segment-balanced objective.

Token-level supervision under the current policy. Following on-policy distillation (Agarwal et al., 2024; Gu et al., 2024), the current student generates a complete trajectory $y = (y_1, \ldots, y_T)$, where $y_t \sim \pi_\theta(\cdot \mid x_S, y_{<t})$. Teacher and student then score identical completion token IDs at the same causal prefixes:

$$p_t^S(v) = \pi_\theta(v \mid x_S, y_{<t}), \qquad p_t^T(v) = \pi_T(v \mid x_T, y_{<t}), \quad v \in \mathcal{V}. \tag{2}$$

Here $\mathcal{V}$ is the shared vocabulary; student replay is differentiable and teacher scoring is gradient-free. The teacher's top-K entries define a common support $\mathcal{K}_t$. We retain probabilities normalized over the full vocabulary and merge the remaining mass into tails τ_t^S, τ_t^T. The position-wise objective is reverse KL from student to teacher:

$$\ell_t = \sum_{v \in \mathcal{K}_t} p_t^S(v) \log \frac{p_t^S(v)}{p_t^T(v)} + \tau_t^S \log \frac{\tau_t^S}{\tau_t^T}. \tag{3}$$

The main configuration fixes $K = 128$ and clips each position's loss to $\hat{\ell}_t = \mathrm{clip}(\ell_t, 0, 2)$. BB-OPD preserves this token-level objective and changes only its aggregation weights.

Reasoning units with local boundary uncertainty. A process mask retains the original indices of positions to aggregate, $\Omega = (u_1, \ldots, u_N)$. The main configuration uses frozen Qwen3-VL-32B-Instruct as an independent text segmenter to merge adjacent micro-clauses into ordered atomic reasoning units, such as visual observations, actions or states, spatiotemporal relations, causal relations, option elimination, and intermediate inferences. It predicts grouping boundaries without rewriting the student trajectory. Validation of micro-clause indices and group IDs yields M contiguous units with boundaries $0 = b_0 < b_1 < \cdots < b_M = N$.

Natural-language actions, outcomes, and explanations rarely admit a unique segmentation. Treating one post-hoc partition as exact would therefore make supervision weights depend on incidental boundary placement. We sample an integer offset around each internal boundary, setting its range to 10% of the shorter adjacent original segment and constraining the resulting boundaries $\tilde{b}_j$ to remain ordered and nonempty. The perturbed unit j is

$$\widetilde{S}_j = \{u_n \mid \tilde{b}_{j-1} < n \le \tilde{b}_j\}, \qquad j = 1, \ldots, M. \tag{4}$$

With $\tilde{b}_0 = 0$ and $\tilde{b}_M = N$, the units still cover all of Ω. This operation neither reorders, removes, nor compresses text, and it leaves the position-wise KL in Eq. 2 unchanged; only the aggregation unit assigned to neighboring tokens may change.

Boundary-balanced objective. For a batch of B valid rollouts, let trajectory i contain M_i perturbed segments $\widetilde{S}_{ij}$. We first average losses within each segment, weight segments equally within each trajectory, and finally average over rollouts:

$$\mu_{ij} = \frac{1}{|\widetilde{S}_{ij}|} \sum_{t \in \widetilde{S}_{ij}} \hat{\ell}_{it}, \qquad \mathcal{L} = \frac{\lambda}{B} \sum_{i=1}^{B} \frac{1}{M_i} \sum_{j=1}^{M_i} \mu_{ij}, \quad \lambda = 0.25. \tag{5}$$

Each token in unit j receives weight $1/(M_i|\widetilde{S}_{ij}|)$, while every unit has total weight $1/M_i$. Semantic grouping provides local structure, boundary perturbation changes where length normalization acts, and equal unit weighting removes the proportional dependence of total unit weight on expression length. The objective neither assumes that shorter units are more important nor treats segmenter outputs as importance annotations.

3.4 Training and Deployment

Each training iteration comprises on-policy sampling, same-trajectory teacher–student scoring, and BB-OPD aggregation. The student samples a trajectory from x_S, and the teacher replays the same text prefixes under x_T, which additionally contains VEM evidence. Equation 3 produces position-wise supervision, which Eq. 5 aggregates into the training objective. VEM evidence is constructed offline, and the segmenter supplies gradient-free boundary metadata. Teacher parameters, distributions, supports, and grouping boundaries remain fixed; gradients flow only through student probabilities to update θ.

At deployment, only the student is retained. It answers directly from the global video, question, and options $x_S = (V_G, q, o)$, without reading an evidence sheet or invoking the teacher, evidence builder, or segmentation service. The two training-time mechanisms therefore add neither deployment inputs nor auxiliary modules to the student.

Table 1: Results on six public video benchmarks (%).

Model	VideoMMMU	MMVU	VSI	MV.	TempC.	Video-MME
LongVA-7B	23.90	–	29.20	–	56.90	52.60
VILA-1.5-40B	34.00	–	31.20	–	–	60.10
LLaVA-OneVision-7B	33.80	49.20	32.40	56.70	–	58.20
Qwen2.5-VL-7B	47.40	61.30	31.80	59.40	69.20	52.80
Video-R1-7B (16f)	49.80	64.20	34.60	62.70	72.60	57.40
Video-R1-7B (32f)	52.30	63.80	35.80	63.90	73.20	59.30
VideoRFT-7B	51.10	68.50	36.80	62.10	73.70	59.80
STRIVE (32f)ST	56.50	69.60	54.40	–	78.30	62.10
Qwen3-VL-8B (16f)	64.85	68.32	40.37	63.42	**74.48**	56.48
ViB-OPD-8B (16f)	**65.42**	**69.60**	**41.29**	**63.71**	74.36	**58.67**
Qwen3-VL-8B (32f)	66.62	70.88	44.48	63.39	75.34	61.70
ViB-OPD-8B (32f)	**67.01**	**72.00**	**45.17**	**64.66**	**76.26**	**63.52**
Qwen3.5-9B (16f)	66.33	72.48	39.22	63.05	74.92	59.59
ViB-OPD-9B (16f)	**68.15**	**74.24**	**41.73**	**65.26**	**76.66**	**60.48**
Qwen3.5-9B (32f)	67.01	72.16	41.26	64.00	75.23	63.59
ViB-OPD-9B (32f)	**69.06**	**72.80**	**42.60**	**65.05**	**76.26**	**65.67**

4 Experiments

4.1 Experimental Setup

Training configuration. We conduct our experiments using a total of 16 NVIDIA H200 GPUs. We use a Qwen3-VL-8B student and a frozen Qwen3-VL-32B teacher. The training set contains 1,000 unique video–question records (200/400/400 short/medium/long) with VEM contact sheets. Training and evaluation indices are constructed separately, without Video-MME training examples. Each batch has four questions and two rollouts per question across four training ranks. Training lasts 200 steps with a 3×10^{-7} learning rate, a 10-step warmup followed by a constant schedule, and a 512-token generation limit. Except in the top-K ablation, we use teacher top-128 plus tail. All groups use a per-token KL cap of 2 and coefficient 0.25, and both models score the same current student trajectory.

Evaluation benchmarks. We evaluate on six public video benchmarks: VSI-Bench (Yang et al., 2024), VideoMMMU (Hu et al., 2025), MMVU (Zhao et al., 2025), MVBench (Li et al., 2023), TempCompass (Liu et al., 2024), and Video-MME (Fu et al., 2024). The first three emphasize video reasoning, while the latter three cover both perception and reasoning to assess overall video understanding. We use the multiple-choice subsets of MMVU and VideoMMMU, a multiple-choice-compatible setting for TempCompass, the debiased test set of VSI-Bench, and the full Video-MME test set without subtitles.Our reevaluations consistently use greedy decoding with temperature 0, top-p 1, and seed 42.

4.2 Main Results

Matched backbone and inference budget. With backbone, visual budget, and decoding protocol held identical, ViB-OPD consistently improves the 8B student on all six benchmarks. Against the base model at the same budget, ViB-OPD improves every benchmark in Table 1 by 1.04 points on average: Video-MME rises from 61.70% to 63.52%, MMVU from 70.88% to 72.00%, and MVBench from 63.39% to 64.66%. The three reasoning-oriented and three perception-inclusive benchmarks improve together, so the gain is not confined to a particular task composition.

Cross-family replication. The gain holds after changing the teacher–student family. We apply the same procedure and hyperparameters—including the 10% perturbation scale and $K = 128$—to a Qwen3.5-9B student and a frozen Qwen3.5-27B teacher, without re-searching any configuration for the new family. All six benchmarks improve at both 16 and 32 frames, with average gains of 1.82 and 1.37 points; on Video-MME the 32-

frame model rises from 63.59% to 65.67%. The gain therefore follows from how the training signal is organized rather than from tuning to one backbone.

Gains under a reduced frame budget. The gain persists when the visual budget is halved, and training-time evidence complements inference-time coverage. At 16 frames, ViB-OPD improves the Qwen3-VL-8B base model on five of six benchmarks by 0.86 points on average; Video-MME rises from 56.48% to 58.67%, while TempCompass decreases slightly by 0.12 points. The 16-frame score remains below the 63.52% obtained with 32 frames, showing that privileged evidence improves how efficiently a limited visual input is used: it acts on the supervision signal, inference-time sampling on what is visible, so the two compose rather than substitute.

Overall and duration-specific Video-MME results. With the same visual teacher, the full method improves accuracy across all three video-duration subsets, with the largest gain on Long. The three aggregation variants using the same privileged-visual teacher in Table 2 share the training data, on-policy sampling, and token-level reverse KL. The full method reaches 63.52% Overall, exceeding token averaging and length-matched random balancing by 1.00 and 1.26 percentage points, respectively.

Table 2: Video-MME accuracy (%).

Method	Overall	Short	Medium	Long
Qwen3-VL-8B	61.70	72.67	59.33	53.11
Global + Token	62.07	73.33	59.78	53.11
VEM + Token	62.52	73.89	60.00	53.67
VEM + Random	62.26	73.67	59.44	53.67
ViB-OPD	**63.52**	**74.00**	**60.78**	**55.78**

VEM adds teacher-side evidence sheets; Token and Random denote token averaging and length-matched random segment balancing.

The full method outperforms both controls on Short, Medium, and Long. The largest accuracy gains occur on Long, at 2.11 and 2.11 percentage points, respectively.

4.3 Ablation Studies

Table 3: **Video-MME ablations (accuracy, %). (a)** Contributions of VEM and BB-OPD with a shared global-video input. **(b)** Loss aggregation with the VEM teacher fixed. Segment-based variants use within-segment means and equal segment weights. Fixed-10 uses 10-token windows; Random-matched preserves the semantic segment-length multiset.

(a) Component ablation.

VEM	BB-OPD	Overall	Short	Medium	Long
–	–	62.07	73.33	59.78	53.11
✓	–	62.52	73.89	60.00	53.67
–	✓	61.89	73.00	59.78	52.89
✓	✓	**63.52**	**74.00**	**60.78**	**55.78**

(b) Loss aggregation.

Method	Overall	Short	Medium	Long
Token	62.52	73.89	60.00	53.67
Fixed-10	61.59	73.22	58.33	53.22
Punctuation	62.22	73.44	59.67	53.56
Semantic	62.70	73.67	60.56	53.89
Random-matched	62.26	73.67	59.44	53.67
BB-OPD	**63.52**	**74.00**	**60.78**	**55.78**

Component ablation of VEM and BB-OPD. The two components depend on each other: reallocating supervision pays off only when the teacher has adequate visual grounding. Table 3a crosses visual input with loss aggregation. Under token averaging, VEM raises Overall from 62.07% to 62.52%; BB-OPD alone reaches 61.89%, below the baseline; together they reach 63.52% and are best on all four metrics. The joint gain of 1.45 points far exceeds the sum of the individual effects: segment-level allocation is only as good as the distribution it redistributes, and reweighting an under-grounded teacher amplifies noise where it is weakest.

Visual input configurations for VEM. With the teacher's visual budget held identical, question-relevant frame selection still yields a gain. The left group of Figure 2 varies input organization: global, evidence-only, and mixed views reach 62.26%, 62.33%, and 62.41%. The right group fixes the 32-frame global video and an identically formatted 16-frame sheet, holding the teacher-side visual budget constant and changing only the selection rule: uniform, random, and VEM reach 62.59%, 62.63%, and 63.52%. VEM leads both by 0.93 and 0.89 points, while uniform and random differ by only 0.04, so the gain comes from question relevance rather than from the number of supplementary frames.

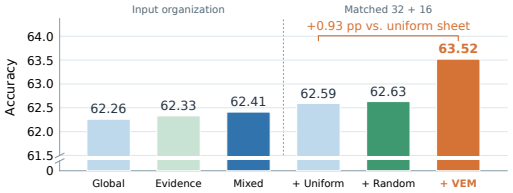

Figure 2: Visual-input comparison for VEM. The left group compares input organization with unmatched visual budgets; the right group fixes a 32-frame global video and a 16-frame sheet while changing only the sheet source. Bar labels report Overall accuracy on full Video-MME (%).

Loss aggregation in BB-OPD. Combining semantic grouping with local boundary perturbation yields the best aggregation result. Table 3b fixes the privileged-visual teacher and token-level reverse KL; every method except token averaging averages within segments and then weights segments equally. Fixed-window, punctuation, length-matched random, and semantic grouping achieve 61.59%, 62.22%, 62.26%, and 62.70% Overall, respectively. Semantic grouping performs best among the fixed schemes, improving on token averaging by 0.18 percentage points. Adding local boundary perturbation to the same semantic initialization further raises Overall to 63.52% and Long to 55.78%, gains of 0.82 and 1.89 points over fixed semantic grouping. These results support the combined design of BB-OPD: semantic units organize segment-level supervision, while local boundary variation adjusts the partitions and token weights within their neighborhoods.

Boundary perturbation scale in BB-OPD. Overall responds to the perturbation scale with a single peak, optimal at 10%. Figure 3(a) fixes code, semantic initialization, data, training steps, K, and seed protocol, and varies only the scale: 1%, 3%, 5%, 10%, and 15% reach 62.96%, 63.00%, 63.15%, 63.52%, and 62.52%, with length-matched random grouping as an independent reference. The response declines on both sides of the operating point, consistent with the mechanism: too small a scale approaches a fixed segmentation, while too large a scale crosses semantic boundaries and merges unrelated tokens into one weighting unit.

Teacher support set. $K = 128$ performs best among the tested support sizes. Figure 3(b) compares Video-MME Overall accuracy across teacher support sizes, retaining tail probability mass in every configuration. Accuracy peaks at $K = 128$, and increasing K further yields no additional gain. We therefore use $K = 128$ in the main experiments and all other ablations.

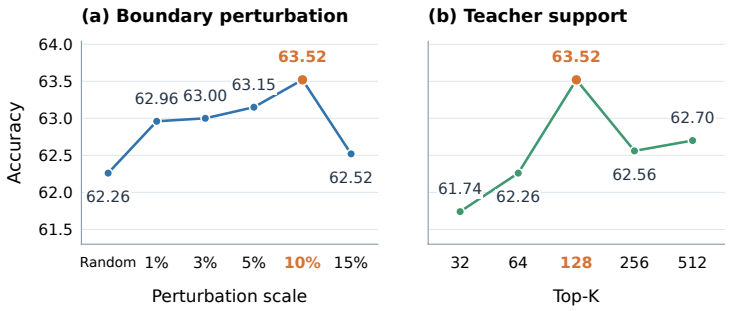

Figure 3: Video-MME ablations (%). **(a)** BB-OPD scales: 1%, 3%, 5%, 10%, 15%; Random is an independent length-matched reference. **(b)** Teacher top-K plus tail.

4.4 Visualizations

Qualitative comparisons on public videos. Figure 4 contrasts student outputs before and after joint VEM and BB-OPD training on visual counting and event succession. Both cases are drawn from questions the base model answers incorrectly and the full method answers correctly; each shows six frames from a 32-frame preview with verbatim archived excerpts, and both students use the same 32-frame inference budget with no teacher at inference time. The two question types—spatial enumeration and temporal ordering—both require evidence to accumulate across intermediate steps rather than a prior over final answers.

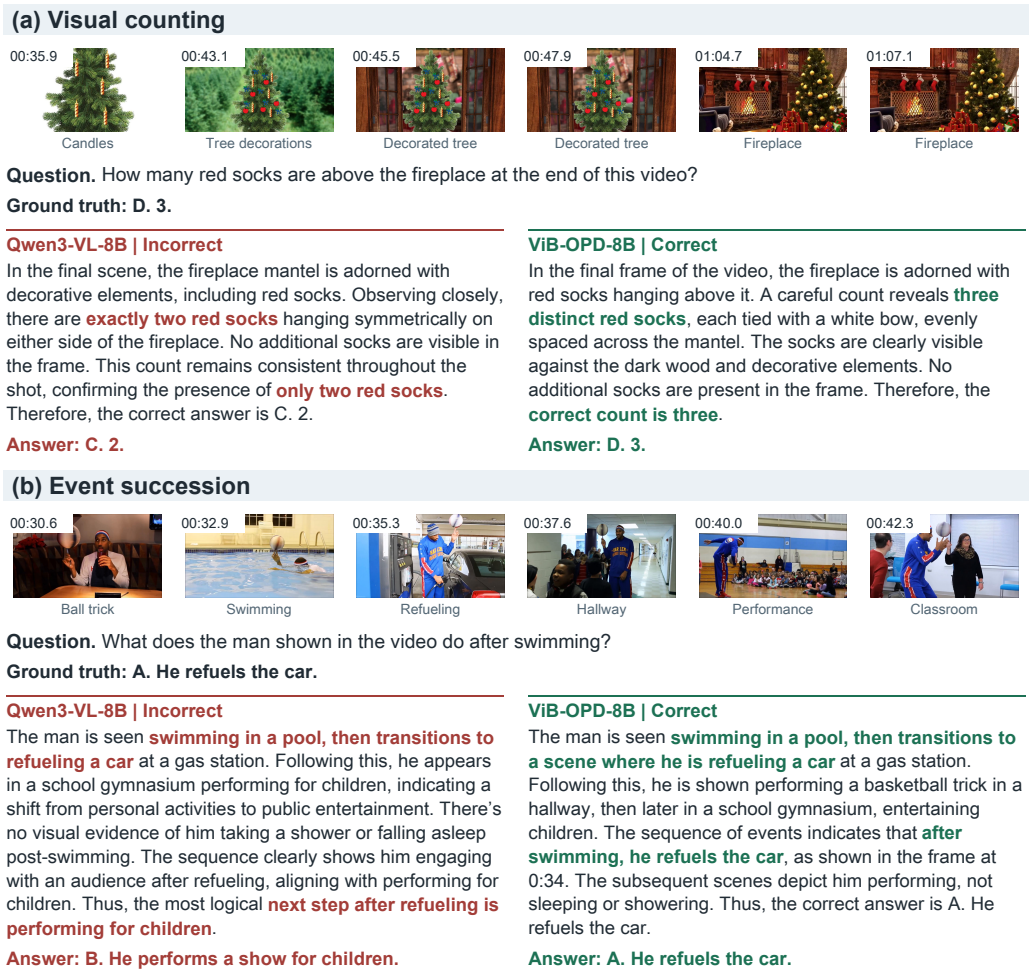

Figure 4: **Video-MME qualitative comparisons.** Both models use 32 frames; six are shown per case. **(a)** Visual counting. **(b)** Event succession. Cases are selected by outcome and excerpts are verbatim.

5 Discussion and Conclusion

We present ViB-OPD, which combines training-only Visual Evidence Mining with Boundary-Balanced OPD. Question-relevant evidence enriches teacher supervision, while semantic grouping, local boundary perturbation, and equal segment weighting organize its allocation. The method preserves student trajectories and token-level reverse KL; inference uses only the student with its original visual input. Across six public benchmarks and two teacher–student families, it improves student performance. On Video-MME, it reaches 63.52%, exceeding token-averaged OPD with the same teacher by 1.00 percentage points.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, et al. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024. URL https://arxiv.org/abs/2406.07476.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, et al. Video-R1: Reinforcing video reasoning in MLLMs. *arXiv preprint arXiv:2503.21776*, 2025. URL https://arxiv.org/abs/2503.21776.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. URL https://arxiv.org/abs/2405.21075.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Xiaoyang Guo, Guoping Luo, Jusheng Zhang, et al. Token-budget distillation: Transferring full-token semantics to compressed video vision-language models. *arXiv preprint arXiv:2608.28138*, 2026. URL https://arxiv.org/html/2608.28138v1.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025. URL https://arxiv.org/abs/2501.13826.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026. URL https://arxiv.org/abs/2603.07079.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024. URL https://arxiv.org/abs/2402.03898.

Ji Soo Lee, Jinyoung Park, Seohyun Lee, Jongha Kim, Joonmyung Choi, et al. Reason in the words you speak: Idiolectal paraphrasing off-policy traces for reasoning distillation in VideoLLMs. *arXiv preprint arXiv:2608.26684*, 2026. URL https://arxiv.org/html/2608.26684v1.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, et al. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. URL https://arxiv.org/abs/2408.03326.

Kunchang Li, Yali Wang, Yinan He, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*, 2023. URL https://arxiv.org/abs/2311.17005.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. URL https://arxiv.org/abs/2311.10122.

Yuanxin Liu et al. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. URL https://arxiv.org/abs/2403.00476.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015. URL https://arxiv.org/abs/1511.03643.

Chuanyu Qin, Chenxu Yang, Qingyi Si, et al. EasyVideoR1: Easier RL for video understanding. *arXiv preprint arXiv:2604.16893*, 2026. URL https://arxiv.org/html/2604.16893v1.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. URL https://arxiv.org/abs/2211.14275.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16(61):2023–2049, 2015. URL https://www.jmlr.org/papers/v16/vapnik15b.html.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*, 2023. URL https://arxiv.org/abs/2312.08935.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. VideoRFT: Incentivizing video reasoning capability in MLLMs via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025. URL https://arxiv.org/abs/2505.12434.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024. URL https://arxiv.org/abs/2407.15754.

Jingpei Wu, Xiao Han, Weixiang Shen, Boer Zhang, Zifeng Ding, and Volker Tresp. ProcessThinker: Enhancing multi-modal large language models reasoning via rollout-based process reward. *arXiv preprint arXiv:2606.11209*, 2026. URL https://arxiv.org/html/2606.11209v1. ICLR Workshop on Logical Reasoning.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024. URL https://arxiv.org/abs/2412.14171.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026. URL https://arxiv.org/abs/2602.12275.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. URL https://arxiv.org/abs/2306.02858.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, et al. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. URL https://arxiv.org/abs/2406.16852.

Yilun Zhao, Lujing Xie, Haowei Zhang, et al. Mmvu: Measuring expert-level multi-discipline video understanding. *arXiv preprint arXiv:2501.12380*, 2025. URL https://arxiv.org/abs/2501.12380.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, et al. MLVU: Benchmarking multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. URL https://arxiv.org/abs/2406.04264.